University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

Sampling ÜGK 2017: Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017

Verner, Martin ; Helbling, Laura Alexandra

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-180405
Published Research Report
Published Version

Originally published at:
Verner, Martin; Helbling, Laura Alexandra (2019). Sampling ÜGK 2017: Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017. Zürich: Institut für Bildungsevaluation.

Universität Zürich[UZH]

Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich

Sampling ÜGK 2017

Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017

Martin Verner

Laura Helbling

Zürich, Mai 2019

Anschrift

Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich
Wilfriedstrasse 15
8032 Zürich

Tel.: 043 268 39 62
Fax: 043 268 39 67
www.ibe.uzh.ch

martin.verner@ibe.uzh.ch

Inhaltsverzeichnis

Liste der im Text verwendeten Abkürzungen

BRR:	Balanced Repeated Replication
HarmoS:	Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule
MOS:	Measure of Size
NRA:	Non-Response Adjustment
RN:	Random Number
SI:	Sampling Interval
TCS:	Target Cluster Size
ÜGK:	Überprüfung des Erreichens der Grundkompetenzen

Liste der im Text verwendeten Kantonskürzel

AG:	Kanton Aargau
AI:	Kanton Appenzell Innerrhoden
AR:	Kanton Appenzell Ausserrhoden
BE_d:	Deutschsprachiger Teil des Kantons Bern
BE_f:	Französischsprachiger Teil des Kantons Bern
BL:	Kanton Basel-Landschaft
BS:	Kanton Basel-Stadt
FR_d:	Deutschsprachiger Teil des Kantons Freiburg
FR_f:	Französischsprachiger Teil des Kantons Freiburg
GE:	Kanton Genf
GL:	Kanton Glarus
GR:	Kanton Graubünden
JU:	Kanton Jura
LU:	Kanton Luzern
NE:	Kanton Neuenburg
NW:	Kanton Nidwalden
OW:	Kanton Obwalden
SG:	Kanton Sankt Gallen
SH:	Kanton Schaffhausen
SO:	Kanton Solothurn
SZ:	Kanton Schwyz
TG:	Kanton Thurgau
TI:	Kanton Tessin
UR:	Kanton Uri
VD:	Kanton Waadt
VS_d:	Deutschsprachiger Teil des Kantons Wallis
VS_f:	Französischsprachiger Teil des Kantons Wallis
ZG:	Kanton Zug
ZH:	Kanton Zürich

1 Einleitung

Im Frühjahr 2017 fand die zweite Erhebung zur Überprüfung des Erreichens der Grundkompetenzen (ÜGK) statt. Dabei wurde auf nationaler und kantonaler Ebene analysiert, inwieweit Schülerinnen und Schüler am Ende des 8. Schuljahres die im Rahmen von HarmoS definierten Grundkompetenzen (Nationale Bildungsziele) in der Schulsprache und in der ersten Fremdsprache erreichen.

Landesweit umfasste die interessierende Population knapp 80'000 Schülerinnen und Schüler, die in rund 3'000 Schulen unterrichtet wurden. Da die Überprüfung sämtlicher unterrichteter Schülerinnen und Schüler des 8. Schuljahres mit einem unverhältnismässig hohen Aufwand verbunden gewesen wäre, wurden Schul- oder Schülerstichproben gebildet. Der vorliegende Bericht dokumentiert die eingesetzten Stichprobenverfahren. Da sich Schulsysteme und verfügbare Informationen, die zur Ziehung einer Stichprobe notwendig sind, von Kanton zu Kanton stark unterscheiden, wurden verschiedene Stichprobenverfahren angewendet.

1.1 Ein- und zweistufige Stichprobenverfahren

Die 26 Kantone bzw. 29 sprachregionalen Kantonsteile[1] lassen sich anhand von zwei verschiedenen Stichprobenverfahren in Gruppen einteilen:

- In den Kantonen AI, AR, BE_f, BS, FR_d, JU, NW, GL, SH, OW, UR, VS_d sowie ZG wurde ein Stichprobenverfahren auf Schülerebene eingesetzt. Das heisst, dass alle Schulen, die eine 8. Klasse führen, zur Teilnahme aufgeboten wurden, innerhalb dieser Schulen jedoch ein bestimmter Anteil der Schülerinnen und Schüler ausgewählt wurde. Die Details zur Vorgehensweise bei der Ziehung von Schülerstichproben kann Kapitel 4 entnommen werden. Diese Gruppe wird im späteren Verlauf der Dokumentation als *Kantone mit einstufigen Stichprobenverfahren* bezeichnet.
- In den Kantonen AG, BE_d, BL, FR_f, GE, GR, LU, NE, SG, SO, SZ, TG, TI, VD, VS_f sowie ZH wurde ein zweistufiges Stichprobenverfahren eingesetzt. Das bedeutet, dass nicht sämtliche Schulen dieser Kantone zwecks Erhebung kontaktiert wurden, sondern in einem ersten Schritt ein Stichprobenverfahren zur Ziehung von Schulen zum Einsatz kam. Die Methodik zu diesem Verfahren wird in Kapitel 3 beschrieben. In einem zweiten Schritt wurden analog zu *Kantonen mit einstufigen Stichprobenverfahren* innerhalb der teilnehmenden

[1] Im weiteren Verlauf des vorliegenden Dokuments werden mit dem Begriff «Kanton» sprachregionale Teile der Kantone BE, FR und VS miteinbezogen.

Schulen Schülerinnen und Schüler gezogen, weshalb die Ausführungen in Kapitel 4 für die hier beschriebenen Kantone ebenfalls ihre Gültigkeit haben. Diese Kantone werden fortan als *Kantone mit zweistufigen Stichprobenverfahren* bezeichnet.

1.2 Stichprobenfehler, Klumpen- und Designeffekte

Wird der Anteil Schülerinnen und Schüler, die den definierten Grundkompetenzen nicht genügen, auf Basis von Stichproben geschätzt, dann kann dieser Anteil in Abhängigkeit der gezogenen Schulen bzw. Schülerinnen und Schüler variieren. In anderen Worten: Würde der Anteil mit verschiedenen Stichproben geschätzt, dann wäre mit einer bestimmten Varianz im geschätzten Anteil zu rechnen (Stichprobenvarianz). Die Streuung dieser Anteilsschätzung widerspiegelt den Stichprobenfehler bzw. Standardfehler, mit dem auf Stichproben beruhende Schätzungen behaftet sind. Dieser Fehler wird in erster Linie von der Varianz des Zielmerkmals und vom Stichprobenumfang bestimmt.[2] Je kleiner die Varianz des Merkmals und je grösser die Stichprobe, umso kleiner fällt der Stichprobenfehler aus. Die Methoden zur Berechnung dieser Fehler bei der ÜGK 2017 sind Kapitel 6 zu entnehmen.

Werden mehrstufige Stichprobenverfahren eingesetzt, besteht im Bildungskontext das Problem, dass sich die Individuen bezüglich des zu messenden Merkmals nicht zufällig auf die Auswahleinheiten der ersten Stufe (z.B. Schulen oder Schulklassen) verteilen. So sind sich Schülerinnen und Schüler innerhalb einer Schule in diversen Merkmalen (vor allem bezüglich schulischer Leistungen, wie sie mit der ÜGK erhoben werden) ähnlicher als Schülerinnen und Schüler aus unterschiedlichen Schulen. In *Kantonen mit zweistufigen Stichprobenverfahren* führt dies dazu, dass bei der Ziehung von Schulen relativ leistungshomogene Gruppen in die Stichproben aufgenommen werden. Diese Klumpeneffekte können den oben beschriebenen Stichprobenfehler vergrössern bzw. die Messpräzision verringern. Auch Schulklassen stellen solche Klumpen dar. Da im Rahmen der ÜGK 2017 jedoch keine Schulklassen gezogen wurden (vgl. 4.4), sind hier Klumpeneffekte auf Klassenebene stichprobentheoretisch irrelevant.

---

[2] Im vorliegenden Fall handelt es sich beim Zielmerkmal um den Anteil Schülerinnen und Schüler, die den Grundkompetenzen nicht genügen. Wie der folgenden Formel entnommen werden kann, wird dabei die Varianz dann maximal, wenn sich der Anteil 0.5 nähert (p entspricht dem zu schätzenden Anteil, n dem Stichprobenumfang und σ dem Standardfehler; Kauermann & Küchenhoff, 2011):

$$\sigma_n = \sqrt{\frac{p(1-p)}{n}}$$

Bei mehreren in Frage kommenden Stichprobendesigns gilt dasjenige als effizienter, das bei gleichbleibendem Stichprobenumfang den tieferen Stichprobenfehler aufweist. Diese Effizienz wird in der Regel mit dem Designeffekt quantifiziert, der die Stichprobenvarianz eines zu beurteilenden Designs ins Verhältnis zur erwarteten Stichprobenvarianz einer einfachen Zufallsstichprobe setzt. So lässt sich auch die Auswirkung von merkmalshomogenen Gruppen auf den Stichprobenfehler bei mehrstufigen Stichprobenverfahren mithilfe des Designeffekts beziffern. In diesem Fall setzt er sich aus der Grösse der Klumpen sowie einem Mass für die Homogenität des Merkmals zusammen.[3] Um den Stichprobenfehler bei komplexeren (zweistufigen) Stichprobenverfahren a priori zu schätzen, wird deshalb die Quadratwurzel des Designeffekts mit dem bei einer einfachen Zufallsstichprobe zu erwartenden Stichprobenfehler multipliziert (Kish, 1965).

1.3 Stichprobenumfänge

Bei der Erarbeitung des Stichprobendesigns wurde von der Vorgabe ausgegangen, dass die maximale Gesamtstichprobengrösse von 21'000 Schülerinnen und Schülern nicht überschritten werden sollte. Darüber hinaus war es das Ziel, eine kantonal vergleichbare Schätzpräzision zu erreichen, während Auswertungen auf nationaler Ebene in möglichst kleinen Standardfehlern resultieren. Um einen entsprechenden Kompromiss zwischen maximaler Schätzpräzision auf kantonaler und nationaler Ebene zu gewährleisten, wurde darauf verzichtet, die Stichprobengrösse gleichmässig auf die 29 Kantonsteile aufzugliedern. Damit möglichst präzise Ergebnisse auf kantonaler Ebene erzielt und kantonsspezifische Stichprobenverfahren ausgearbeitet werden konnten, wurde die Population in 29 Schichten aufgeteilt, die jeweils einem expliziten Stratum mit separatem Ziehungsdesign entsprachen (vgl. 3.1).

Kantone mit einstufigen Stichprobenverfahren zeichneten sich dadurch aus, dass die Schülerschaft im 8. Schuljahr auf verhältnismässig wenige Schulen verteilt war. Die Bildung einer Schulstichprobe in diesen Kantonen hätte dazu geführt, dass zur Erreichung des anvisierten Stichprobenumfangs ein relativ hoher Anteil der Schülerinnen und Schüler einer gezogenen Schule am Test hätte teilnehmen müssen. Der durch die Ziehung von Schulen entstehende Klumpeneffekt hätte die Schätzpräzision jedoch derart beeinträchtigt, dass es sich stattdessen anbot, sämtliche Schulen der entsprechenden Kantone zu berücksichtigen (Vollerhebung auf Ebene der Schulen) und innerhalb der einzelnen Schulen jeweils eine Zufallsstichprobe von Schülerinnen und

[3] Der Designeffekt ($deff$) aufgrund komplexer Stichprobendesigns ist abhängig von der mittleren Klumpengrösse (b) und dem Intraklassenkoeffizienten (ρ) des entsprechenden Merkmals (vgl. Kish, 1965): $deff = 1 + \rho(b - 1)$

Schülern zu ziehen. Der Richtwert für Stichprobengrössen für *Kantone mit einstufigen Stichprobenverfahren* betrug – sofern der Populationsumfang dies ermöglichte – 600 Schülerinnen und Schüler. Da abwesende, ausgeschlossene oder verweigernde Schülerinnen und Schüler nicht einberechnet werden und die definitiven Stichprobenumfänge zusätzlich von der Varianz der Schülerbestände zwischen den Schulen abhängig waren (in *Kantonen mit einstufigen Stichprobenverfahren* wurden in verhältnismässig grossen Schulen mehrere Testsitzungen durchgeführt), weichen die in Tabelle 1.1 dargestellten Zahlen vom Richtwert ab.

Tabelle 1.1: Anzahl getesteter Schülerinnen und Schüler sowie Populationsumfänge getrennt nach Kanton für *Kantone mit einstufigen Stichprobenverfahren*

Kanton	Populationsumfang	Realisierte Stichprobe
AI	141	139
AR	468	441
BE_f	795	527
BS	1'418	628
FR_d	758	513
GL	339	254
JU	799	595
NW	340	273
OW	351	246
SH	646	534
UR	327	290
VS_d	745	572
ZG	1'034	577
Total	*8'161*	*5'589*

Anmerkungen: Bei den Stichprobengrössen handelt es sich um die Anzahl tatsächlich am Test teilnehmender Schülerinnen und Schüler. Die Populationsumfänge beruhen auf der jeweiligen kantonalen Summe der Schülergewichte und beziehen sich demnach auf die *ÜGK-Population* (vgl. Kapitel 2). Abhängig von Ausschlüssen und der Güte der für die Ziehung von Schulen verwendeten Schülerbestandslisten, können diese Zahlen von den tatsächlichen, kantonalen Populationsumfängen abweichen.

Um Klumpeneffekten entgegenzuwirken, wurden in *Kantonen mit zweistufigen Stichprobenverfahren* verhältnismässig mehr Schülerinnen und Schüler gezogen. Die entsprechenden Stichproben- sowie Populationsumfänge sind in Tabelle 1.2 aufgeführt. Der ursprüngliche Richtwert entsprach einem Stichprobenumfang von 900 bis 1'000 Schülerinnen und Schülern pro Kanton. Um einem aufgrund variierender Gewichte wachsenden Stichprobenfehler (vgl. hierzu Kish, 1992) auf nationaler Ebene entgegenzuwirken, wurde der Stichprobenumfang im bevölkerungsreichsten Kanton (ZH) deutlich erhöht. Die in Tabelle 1.2 dargestellten Stichprobenumfänge waren ebenfalls abhängig von der Varianz der Schülerbestände zwischen Schulen eines Kantons und

beinhalten keine abwesenden, ausgeschlossenen oder verweigernden Schülerinnen und Schüler.

Tabelle 1.2: Anzahl getesteter Schülerinnen und Schüler sowie Populationsumfänge getrennt nach Kanton für *Kantone mit zweistufigen Stichprobenverfahren*

Kanton	Populationsumfang	Realisierte Stichprobe
AG	6'142	906
BE_d	8'173	940
BL	2'385	865
FR_f	2'645	955
GE	4'610	892
GR	1'315	820
LU	3'798	940
NE	1'835	643
SG	4'584	951
SO	2'275	912
SZ	1'311	654
TG	2'800	927
TI	3'135	744
VD	7'669	928
VS_f	2'425	944
ZH	13'411	1'567
Total	*68'513*	*14'588*

Anmerkungen: Bei den Stichprobengrössen handelt es sich um die Anzahl tatsächlich am Test teilnehmender Schülerinnen und Schüler. Die Populationsumfänge beruhen auf der jeweiligen kantonalen Summe der Schülergewichte und beziehen sich demnach auf die *ÜGK-Population* (vgl. Kapitel 2). Abhängig von Ausschlüssen und der Güte der für die Ziehung von Schulen verwendeten Schülerbestandslisten, können diese Zahlen von den tatsächlichen, kantonalen Populationsumfängen abweichen.

1.4 Stichprobengewichte

Für jede an der ÜGK 2017 teilnehmende Schülerin und jeden teilnehmenden Schüler wurde ein Stichprobengewicht berechnet. Die Schülergewichte sind ein Mass für die Anzahl Schülerinnen und Schüler in der Population, die durch die entsprechenden Schülerinnen und Schüler in der Stichprobe repräsentiert werden. Mit anderen Worten: Die Summe der Gewichte sämtlicher an der ÜGK 2017 teilnehmender Schülerinnen und Schüler entspricht näherungsweise dem Umfang der *ÜGK-Population* (vgl. Kapitel 2). Die Gewichte kompensieren primär die unterschiedlichen individuellen Auswahlwahrscheinlichkeiten der einzelnen Schülerinnen und Schüler, können jedoch abhängig von Absenzen und Verweigerungen nachträglich korrigiert worden sein. Einzelheiten zur Gewichtung werden in Kapitel 5 berichtet.

2 Population

In diesem Kapitel wird die *ÜGK-Population* definiert bzw. diejenige Menge von Schülerinnen und Schülern beschrieben, auf welche die ÜGK 2017 Rückschlüsse erlaubt. Dies beinhaltet auch die Erläuterung diverser Ausschlusspraktiken auf Schul- sowie Schülerebene (Abweichungen von der erwünschten Population).

Bei landesweiten oder internationalen Erhebungen der Schulleistung werden in der Regel Alterskohorten oder ein interessierendes Schuljahr definiert. Da bei der ÜGK 2017 die Schülerinnen und Schüler am Ende des 8. Schuljahres im Mittelpunkt standen, bot sich eine auf dem Schuljahr beruhende Populationsdefinition geradezu an. Verglichen mit der Methode der Alterskohorte sind die Festlegung einer Population, das Ziehen einer Stichprobe sowie die Durchführung der Erhebung bei einer durch das Schuljahr definierten Population vergleichsweise einfach (Rust, 2014).

Dementsprechend umfasste die *erwünschte Population* der ÜGK 2017 sämtliche Schülerinnen und Schüler, die in einer nach Schweizer Recht organisierten Schule im 8. Schuljahr unterrichtet wurden. Gemäss dieser Definition waren private Schulen (unabhängig vom Subventionsgrad) grundsätzlich auch Teil der Population. Lediglich Schulen, die auf Basis von ausländischen Programmen oder in keiner Landessprache unterrichteten, waren in dieser Definition nicht berücksichtigt. In der Schweiz trifft dies auf eine äusserst kleine Anzahl internationaler Schulen zu, die ausschliesslich englisch- oder japanischsprachigen Unterricht anbieten.

Es war ein erklärtes Ziel der ÜGK, dass die Stichprobe die so definierte *erwünschte Population* möglichst lückenlos abdeckt. Dennoch waren Ausschlüsse sowohl auf Schulebene als auch auf der Ebene der Schülerinnen und Schüler – vorwiegend aus erhebungspraktischen Gründen – unvermeidbar. Die tatsächlich untersuchte *ÜGK-Population*, auf die sich die gewichteten Ergebnisse der ÜGK 2017 beziehen, umfasste aufgrund der im nächsten Abschnitt erläuterten Ausschlüsse weniger Schülerinnen und Schüler als die *erwünschte Population*.

2.1 Ausschlüsse auf Schulebene

Bei der ÜGK 2017 wurden auf Schulebene Sonderschulen aus diversen Gründen ausgeblendet:

- Die grosse Mehrheit der in Sonderschulen unterrichteten Schülerinnen und Schüler kann nicht einem Schuljahr zugeordnet werden, was die Bildung einer klar definierten Zielgruppe deutlich erschwert.
- Den wenigsten Kantonen stehen Informationen über die Häufigkeit bestimmter Behinderungsformen bzw. Lern- und Verhaltensschwierigkeiten an einzelnen

Sonderschulen zur Verfügung. Ohne diese Informationen lässt sich nicht feststellen, an welchen Sonderschulen eine Testdurchführung im Sinne einer objektiven und vor allem zumutbaren Erhebung überhaupt möglich ist.

- Die Testaufgaben wurden nicht im Hinblick auf Sonderschulen entwickelt. Zahlreiche in Sonderschulen unterrichtete Schülerinnen und Schüler könnten die Aufgaben ohne fremde Unterstützung nicht bearbeiten.

Da sich an Sonderschulen unterrichtete Schülerinnen und Schüler nicht einem Schuljahr zuordnen lassen, kann nur grob geschätzt werden, wie viele Schülerinnen und Schüler der *erwünschten Population* ÜGK 2017 durch diesen Ausschluss betroffen waren. Hierbei handelt es sich also um in Sonderschulen unterrichtete Schülerinnen und Schüler, die in einer 8. Klasse gewesen wären, wenn sie eine Regelschule besucht hätten. Mit der Annahme, dass der Anteil Sonderschülerinnen und Sonderschüler zwischen einer bestimmten Alterskohorte (Jahrgang) und einer Klassenstufe vergleichbar ist, kann davon ausgegangen werden, dass ca. 1.6 Prozent der *erwünschten Population* in Sonderschulen unterrichtet wurden.[4] Die geschätzten Anteile für die einzelnen Kantone variieren relativ stark und werden in der zweitletzten Spalte in Tabelle A.1 (siehe Anhang A) dargestellt.

2.2 Ausschlüsse auf Schülerebene

Innerhalb der zur Teilnahme bestimmten Schulen hatten die jeweiligen Schulleitungen bzw. Lehrpersonen die Möglichkeit, einzelne Schülerinnen und Schüler, beruhend auf den folgenden Kriterien, von der Erhebung zu dispensieren:

- Kognitiv beeinträchtigte Schülerinnen und Schüler, deren Beeinträchtigung gemäss zuständigen Lehrpersonen eine valide Testdurchführung verunmöglichte, wurden ausgeschlossen. Hierzu gehören in Regelschulen unterrichtete Schülerinnen und Schüler, die auf emotionaler oder kognitiver Ebene den allgemeinen Anweisungen der Tests nicht folgen können und für die dementsprechend eine Testdurchführung als nicht zumutbar eingestuft wurde.
- Funktional beeinträchtigte Schülerinnen und Schüler, deren körperliche Beeinträchtigung die Validität der Ergebnisse einschränken könnte, wurden ebenfalls nicht einbezogen. Hierbei handelt es sich hauptsächlich um Schülerinnen und Schüler mit Körper- oder Sehbehinderungen.
- Schliesslich wurden auch Schülerinnen und Schüler mit sehr schlechten Kenntnissen der Testsprache ausgeschlossen. Notwendige

[4] Vgl. die Statistik der Lernenden Schuljahr 2016/17 des Bundesamts für Statistik.

Ausschlusskriterien waren, dass (1) die Muttersprache nicht der Testsprache entsprach, (2) die sprachlichen Schulleistungen deutlich eingeschränkt waren und (3) die Schülerin oder der Schüler weniger als ein Jahr in der Testsprache unterrichtet wurde.

Die Schulen wurden ausdrücklich darauf hingewiesen, dass schlechte Schulleistungen oder disziplinarische Probleme keinen Ausschlussgrund darstellen. Dennoch variieren sowohl die Anzahl der als auch die Gründe für Ausschlüsse innerhalb Schulen zwischen den Kantonen relativ stark.

Es soll deutlich darauf hingewiesen werden, dass Tabelle A.1 lediglich der Schätzung kantonaler Ausschlussquoten dient. Die Zahlen können nicht zur Berechnung der Häufigkeit einzelner Behinderungsformen verwendet werden, da die grosse Mehrheit der Schülerinnen und Schüler mit besonderen Lernbedürfnissen an den Tests teilgenommen hat. Darüber hinaus wurden den Tests ferngebliebene Schülerinnen und Schüler, die nicht explizit von der Schulleitung bzw. der Lehrperson als ausgeschlossen kommuniziert worden waren, nicht ausgeschlossen, sondern als «abwesend» vermerkt und dementsprechend bei der Korrektur der Stichprobengewichte berücksichtigt (vgl. 5.4).

3 Schulstichproben

Die vergleichsweise hohe Anzahl Schulen in den Kantonen AG, BE_d, BL, FR_f, GE, GR, LU, NE, SG, SO, SZ, TG, TI, VD, VS_f sowie ZH bedingte die Ziehung von Schulstichproben. Dazu wurde beim Bundesamt für Statistik (BfS) eine Liste sämtlicher Schulen, die Schülerinnen und Schüler im 8. Schuljahr unterrichten, eingefordert. Für sämtliche Schulen enthielt diese Liste – nebst Schulnamen und Adresse – Angaben zu Trägerschaft sowie zur Anzahl unterrichteter Schülerinnen und Schüler im 8. Schuljahr. Schliesslich wurde die Liste der wählbaren Schulen mithilfe der kantonalen Bildungsstatistiken teilweise ergänzt bzw. aktualisiert.

3.1 Stratifizierung der Liste wählbarer Schulen

Um den Einsatz unterschiedlicher Stichprobenverfahren in den verschiedenen Kantonen zu ermöglichen, die Effizienz dieser Verfahren zu erhöhen und um sicherzustellen, dass alle Teile der Population adäquat in der Stichprobe vertreten waren, wurde in *Kantonen mit zweistufigen Stichprobenverfahren* die Liste wählbarer Schulen vor dem Ziehungsprozess nach bestimmten Merkmalen geschichtet und sortiert.

Analog zu den im Rahmen von PISA verwendeten Stichprobenverfahren (OECD, 2017) wurde bei der Vorbereitung der Schullisten sowohl auf explizite als auch auf implizite Stratifizierungsmethoden zurückgegriffen. Erstere beinhalten die Bildung von Schichten (Strata), die im Verlauf des Stichprobenprozesses unabhängig voneinander behandelt werden (vgl. Rust, 2014, S. 126; Meinck, 2015). So bildet jeder Kanton ein explizites Stratum, was den Einsatz kantonsspezifischer Verfahren ermöglichte. Mit Ausnahme von Sonderklassen werden im 8. Schuljahr keine leistungsabhängigen Schulprogramme differenziert. Aus diesem Grund wurden bei der ÜGK 2017 keine expliziten Strata für Schulen innerhalb der Kantone gebildet.

Die implizite Stratifizierungsmethode bezieht sich auf die Sortierung der Schullisten nach bestimmten Schulmerkmalen. Eine adäquate Sortierung innerhalb der expliziten Strata kann dann zu einer Reduktion des Stichprobenfehlers führen, wenn die Ziehung der Schulen auf eine systematische Art und Weise durchgeführt wird (vgl. Rust, 2014, S. 129). Der Begriff «systematisch» bedeutet in diesem Zusammenhang, dass ein auf einer Zufallszahl beruhendes Ziehungsintervall definiert wird, mit dem die sortierten Listen «durchgezählt» und die entsprechenden «Treffer» als gezogene Einheiten gelten (vgl. 3.2).

Innerhalb eines Kantons wurden die Schulen dementsprechend nach Trägerschaft (öffentliche vs. private Schulen) und – um eine Stichprobe mit ausschliesslich kleinen oder grossen Schulen zu verhindern – nach der geschätzten Anzahl unterrichteter Schülerinnen und Schüler sortiert.

3.2 Systematische Ziehung der Schulen per PPS-Verfahren

Analog zu etablierten, internationalen *Large-Scale Assessments* (z.B. *Trends in International Mathematics and Science Study*, TIMSS; *Progress in International Reading Literacy Study*, PIRLS; *Program of International Student Assessment*, PISA) wurden in *Kantonen mit zweistufigen Stichprobenverfahren* in einem ersten Schritt Schulen proportional zu ihrer Grösse (*Probability Proportional to Size*; *PPS*; z.B. Rust, 2014) gezogen und in einem zweiten Schritt eine bestimmte Anzahl Schülerinnen und Schüler pro Schule zur Teilnahme an der Erhebung aufgeboten. Aus diesem Grund wurde jeder Schule eine *Measure of Size* (*MOS*) zugeordnet, die grundsätzlich der geschätzten Anzahl im 8. Schuljahr unterrichteter Schülerinnen und Schüler entsprach (mit Ausnahme kleiner Schulen, vgl. 3.4). In der grossen Mehrheit der Kantone enthielten die Schullisten Angaben aus älteren Schuljahren. Dementsprechend kann die *MOS* von der tatsächlichen Anzahl unterrichteter Schülerinnen und Schüler abweichen.

Anstatt einer einfachen Zufallsstichprobe wurde ein systematisches Verfahren eingesetzt. Dieses Verfahren beinhaltet die systematische Sortierung der Listen sämtlicher Einheiten in der Population nach bestimmten Merkmalen (vgl. implizite Stratifizierung, 3.1). Grundsätzlich werden dabei Samplingintervalle (*SI*) definiert, die sich aus dem Populationsumfang *N* dividiert durch die erwünschte Stichprobengrösse *n* berechnen. Vor der Ziehung wird eine Startzahl (*RN*) entsprechend einem Zufallswert zwischen 0 und *SI* definiert. Die gezogenen Einheiten in der Liste entsprechen den «Auswahlnummern» *RN*, *RN* + *SI*, *RN* + 2*SI* usw. bis *RN* + (*n* - 1) *SI* (vgl. Rust, 2014, S. 129).

Im konkreten Fall der ÜGK 2017 wurde das systematische Prinzip auf die *PPS*-Ziehung von Schulen in *Kantonen mit zweitstufigen Stichprobenverfahren* übertragen: Da *N* der Summe von *MOS* entsprach (MOS_{tot}), ergab sich für jedes explizite Stratum *SI* aus MOS_{tot} dividiert durch die Anzahl der zu ziehenden Schulen *n*. Schulen, deren *MOS* gleich oder grösser *SI* war, hatten eine Wahrscheinlichkeit von 1, um in die Stichprobe zu gelangen, und wurden deshalb in jeweils separate Strata – *certainty strata* – überführt, bevor MOS_{tot} und *SI* neu berechnet wurden. Dieser iterative Prozess wurde so lange wiederholt, bis die *MOS* aller Schulen kleiner war als *SI*.

Anschliessend wurde für jedes explizite Stratum eine auf vier Kommastellen gerundete Zufallszahl zwischen 0 und 1 (*RN*) generiert. Das Produkt aus *RN* und *SI* entsprach der ersten «Auswahlnummer» jedes Stratums. Die erste zu ziehende Schule war dementsprechend die erstgelistete Schule, deren kumulative *MOS* (MOS_{cum}) gleich oder grösser als die «Auswahlnummer» war. Wurde zu der ersten «Auswahlnummer» *SI* hinzuaddiert (*RN SI* + *SI*), entsprach dies der zweiten «Auswahlnummer». Die dritte «Auswahlnummer» war die Summe aus zweiter «Auswahlnummer» und *SI* (*RN SI* + 2 *SI*) usw. bis zur letzten «Auswahlnummer» *RN SI* + (*n* - 1) *SI*. Die jeweils erstgelisteten Schulen, deren MOS_{cum} gleich oder grösser als die

«Auswahlnummern» war, fielen in die Stichprobe. Diese «Auswahlnummern» wurden unabhängig für jeden Kanton berechnet, indem jeweils neue *RN* generiert wurden.

3.3 Ersatzschulen

Sofern es die Anzahl vorhandener Schulen pro Stratum erlaubte, wurden jeder gezogenen Schule zwei Ersatzschulen zugeordnet. Damit die Ersatzschulen den erstgezogenen Schulen hinsichtlich impliziter Stratifizierungsvariablen (Schulgrösse, Trägerschaft) möglichst ähnlich waren, wurden jeweils diejenigen Schulen als Ersatzschulen bestimmt, die in der stratifizierten Liste unmittelbar vor und nach der gezogenen Schule aufgeführt waren. Vor allem bei grösseren Schulen war diese Methode nicht anwendbar, weil manchmal zwei aufeinanderfolgende Schulen in die Stichprobe fielen. In solchen Fällen wurde teilweise dieselbe Schule als Ersatzschule für mehrere erstgezogene Schulen definiert. Ersatzschulen wurden nur dann zur Teilnahme aufgefordert, wenn die ursprünglich gezogene Schule die Teilnahme verweigerte. Inwieweit Ersatzschulen zum Einsatz kamen, kann Tabelle B.4 im Anhang entnommen werden.

3.4 Umgang mit kleinen Schulen

Während der Bildung der Strata auf Schulebene wurden die Schulen zusätzlich vier Kategorien, beruhend auf der geschätzten Anzahl unterrichteter Schülerinnen und Schüler, zugeteilt. Der Richtwert der zu ziehenden Schülerinnen und Schüler in *Kantonen mit zweistufigen Stichprobenverfahren* betrug 20 (*Target Cluster Size, TCS*; vgl. 4.2). Sämtliche Schulen mit mindestens 20 *(= TCS)* unterrichteten Schülerinnen und Schülern im 8. Schuljahr wurden als *gross* eingestuft. Schulen mit einer Schülerzahl zwischen 10 *(= TCS/2)* und 20 galten als *mittelgross*, solche mit weniger als zehn, aber mindestens drei Schülerinnen und Schülern wurden als *klein* bezeichnet. Die restlichen Schulen wurden als *sehr klein* eingestuft. Sämtliche gezogenen Schulen wurden kontaktiert und über ihre Teilnahme an der ÜGK 2017 informiert. Betrug die Anzahl tatsächlich im 8. Schuljahr unterrichteter Schülerinnen und Schüler jedoch weniger als vier, wurde aus ökonomischen Gründen auf eine Testdurchführung verzichtet. Der Ausfall dieser Schulen wurde auf Schulebene mit Anpassungen der Stichprobengewichte kompensiert (vgl. 5.3).

Enthielt ein Stratum Schulen mit weniger als 20 Schülerinnen und Schülern, bestand das Risiko, dass die Anzahl gezogener Schülerinnen und Schüler dem erwünschten Stichprobenumfang nicht genügt. Darüber hinaus kann eine Schulstichprobe mit zahlreichen kleinen Schulen zu einem unverhältnismässig hohen Aufwand führen, da jeweils Testsitzungen mit äusserst wenigen Schülerinnen und Schülern organisiert werden müssen. Um diese Probleme zu beheben, wurde ein Verfahren angewendet,

das sich an den Vorgehensweisen vergangener PISA-Erhebungen orientiert (OECD, 2017, S. 77).

Bei einer verhältnismässig grossen Anzahl kleiner Schulen innerhalb eines Stratums wurde die Auswahlwahrscheinlichkeit *kleiner* sowie *sehr kleiner* Schulen um die Faktoren 0.5 bzw. 0.25 verringert[5], während gleichzeitig die Auswahl *grosser* Schulen proportional erhöht wurde. Zusammengefasst enthielt dieses Verfahren die folgenden Überprüfungen:

- Wenn der Anteil von in *kleinen* sowie *sehr kleinen* Schulen unterrichteten Schülerinnen und Schülern 1 Prozent oder mehr betrug, wurden diese Schulen unterrepräsentiert und die Anzahl zu ziehender Schulen erhöht.
- Wenn der Anteil von in *kleinen* sowie *sehr kleinen* Schulen unterrichteten Schülerinnen und Schülern unter 1 Prozent lag, der Anteil in *mittelgrossen* Schulen jedoch mindestens 4 Prozent entsprach, wurde lediglich die Anzahl zu ziehender Schulen erhöht. Auf ein *Undersampling kleiner* sowie *sehr kleiner* Schulen wurde in diesem Fall verzichtet.

War keine dieser Bedingungen gegeben, war die Wahrscheinlichkeit gering, dass der relativ kleine Anteil *kleiner* und *sehr kleiner* Schulen die gewünschte Stichprobengrösse in einem relevanten Ausmass reduziert. In diesem Fall wurden weder Schulen unterrepräsentiert noch der Stichprobenumfang angehoben. Die detaillierten Berechnungsschritte dieser Überprüfungen sind in Tabelle B6 im Anhang dargestellt.

[5] Die *MOS* von *mittelgrossen* Schulen betrug stets 20. Dementsprechend war beispielsweise die Auswahlwahrscheinlichkeit einer Schule mit geschätzten 12 Schülerinnen und Schülern dieselbe wie jene einer Schule mit geschätzten 20 unterrichteten Schülerinnen und Schülern. Im Falle einer Unterrepräsentation wurden die *MOS* von *kleinen* sowie *sehr kleinen* Schulen auf 10 (*TCS*/2) bzw. 5 (*TCS*/4) gesetzt.

4 Schülerstichproben

Während in *Kantonen mit zweistufigen Stichprobenverfahren* anhand des in Kapitel 3 beschriebenen Stichprobenverfahrens Schulen zur Teilnahme an der ÜGK 2017 gezogen wurden, nahmen in den restlichen Kantonen sämtliche Schulen an den Erhebungen teil (verweigernde Schulen ausgenommen). In *allen* Kantonen wurde auf Schülerebene ein Stichprobenverfahren angewendet, das in den folgenden Abschnitten beschrieben wird.

4.1 Listen wählbarer Schülerinnen und Schüler

Nachdem die teilnehmenden Schulen bestimmt waren, wurden diese erstmals brieflich kontaktiert. Nebst dem Versand eines allgemeinen Informationsschreibens zur Studie wurden – mit Ausnahme der Kantone GE, NE, TI und VD, bei welchen die Schülerlisten mithilfe eines zentralen, kantonalen Registers erstellt wurden – Listen sämtlicher Schülerinnen und Schüler, die im 8. Schuljahr unterrichtet wurden, eingefordert. Die Schulleitungen wurden gebeten, die Namenslisten mit den folgenden Informationen zu ergänzen:

- Geschlecht
- Geburtsdatum
- Angaben zu allfälligen Lernzielbefreiungen
- Klassenbezeichnung
- Name Klassenlehrperson

Diese Listen bildeten die Grundlage zur Ziehung der teilnehmenden Schülerinnen und Schüler.

4.2 Stichprobenumfänge auf Schülerebene

Vor der Stichprobenziehung auf Schülerebene war es notwendig, die Anzahl der an der Studie teilnehmenden Schülerinnen und Schüler pro Schule zu bestimmen (*Target Cluster Size, TCS*). Zusätzlich zu dem in Abschnitt 1.2 beschriebenen Designeffekt aufgrund komplexer Stichprobendesigns können auch zu stark variierende Stichprobengewichte zu Einbussen in der Schätzpräzision führen (Kish, 1995; Liu, Iannacchione & Byron, 2002; Le, Brick & Kalton, 2002). Aus diesem Grund und weil die Stichprobengewichte umgekehrt proportional zur Auswahlwahrscheinlichkeit sind (vgl. Kapitel 5), wurde innerhalb jedes expliziten Stratums versucht, die Schülerinnen und Schüler mit einer möglichst vergleichbaren Auswahlwahrscheinlichkeit zu ziehen.

In *Kantonen mit zweistufigen Stichprobenverfahren* führte eine konstante *TCS* aufgrund der *PPS*-Selektion (vgl. 3.2) auf Schulebene zu vergleichbaren Stichprobengewichten:

Die Auswahlwahrscheinlichkeiten grosser Schulen waren verhältnismässig hoch, während die Wahrscheinlichkeit einer einzelnen Schülerin bzw. eines einzelnen Schülers, in die schulinterne Stichprobe zu gelangen, bei grossen Schulen relativ gering war (vgl. Rust, 2014, S. 130). Auf theoretischer Ebene ist das Produkt dieser beiden Wahrscheinlichkeiten für alle Schülerinnen und Schüler – unabhängig von der Schulgrösse – identisch, wenn in jeder Schule die gleiche Anzahl Schülerinnen und Schüler in die Stichprobe aufgenommen wird (mit Ausnahme *kleiner* bzw. *sehr kleiner* Schulen; vgl. 3.4). Da die Testadministratorinnen und Testadministratoren mit 20 Tabletcomputern, auf welchen die Tests und Fragebogen dargeboten wurden, ausgestattet waren und um allzu grosse Klumpeneffekte zu vermeiden (vgl. 1.2), wurde eine *TCS* von 20 Schülerinnen und Schülern als adäquat betrachtet.

In *Kantonen mit einstufigen Stichprobenverfahren* hatten alle Schulen dieselbe Wahrscheinlichkeit von $p = 1$, um in die Stichprobe zu gelangen. Um auch hier möglichst identische Schülerauswahlwahrscheinlichkeiten zu erzielen, wurden in einer kleinen Teilmenge verhältnismässig grosser Schulen zwei Testsitzungen durchgeführt und deshalb 40 Schülerinnen und Schüler in die Stichprobe aufgenommen. Dies hatte zur Folge, dass in *Kantonen mit einstufigen Stichprobenverfahren* die Anzahl getesteter Schülerinnen und Schüler pro Schule stärker variierte als in *Kantonen mit zweistufigen Stichprobenverfahren*.

Folgende Aspekte der Schülerstichproben konnten bei der ÜGK 2017 Ursachen für variable Stichprobengewichte darstellen:

- Aufgrund von Schätzungen auf Basis von Schullisten vergangener Schuljahre wich in *Kantonen mit zweistufigen Stichprobenverfahren* die *MOS* teilweise von der tatsächlichen Anzahl unterrichteter Schülerinnen und Schüler ab. Dies bedeutet, dass die Auswahlwahrscheinlichkeit der Schule nicht immer genau proportional zu den Schülerbeständen war. Im Rahmen von PISA werden die Schulgewichte bei extremen Abweichungen zwischen erwarteten und tatsächlichen Schülerbeständen deshalb angepasst (OECD, 2017, S. 118). Die entsprechenden Kriterien trafen bei der ÜGK 2017 jedoch in keinem Fall zu, weshalb die Schulgewichte stets auf Basis der ursprünglichen Auswahlwahrscheinlichkeit belassen wurden.
- Lag in Kantonen mit Schulstichproben die Anzahl unterrichteter Schülerinnen und Schüler unter dem *TCS*-Richtwert (*TCS* = 20), wurden sämtliche Schülerinnen und Schüler zu den Tests aufgeboten bzw. betrug die Auswahlwahrscheinlichkeit auf Schülerebene dann immer $p = 1$. Dementsprechend wurde bei *mittelgrossen* Schulen die *MOS* ebenfalls konstant gehalten (*MOS* = 20). Wurden jedoch *kleine* bzw. *sehr kleine* Schulen unterrepräsentiert (*TCS* = 10 bzw. *TCS* = 5, vgl. 3.4), führte dies zu einer leicht erhöhten Varianz in den Stichprobengewichten.

- In *Kantonen mit einstufigen Stichprobenverfahren* wurden aus organisatorischen Gründen – es wurde versucht, stets sämtliche verfügbaren Testtablets einzusetzen – 20 oder 40 Schülerinnen und Schüler pro Schule in die Stichprobe aufgenommen. Um maximal identische Schülergewichte zu erzielen, müsste jedoch der Anteil gezogener Schülerinnen und Schüler pro Schule mit dem kantonalen Auswahlsatz (Anteil gezogener Schülerinnen und Schüler im Verhältnis zur Population) übereinstimmen. Das bedeutet, dass *in Kantonen mit einstufigen Stichprobenverfahren* maximal belegte Testsitzungen gegenüber nicht-variierenden Schülergewichten priorisiert wurden.
- Nach den Erhebungen war es aufgrund von Verweigerungen und Abwesenheiten teilweise notwendig, die Stichprobengewichte anzupassen. Die entsprechenden Korrekturfaktoren werden in den Abschnitten 5.3 und 5.4 erläutert.

4.3 Stratifizierung auf Schülerebene

Ähnlich wie dies bei der Liste der wählbaren Schulen der Fall war (vgl. 3.1), wurden die Listen wählbarer Schülerinnen und Schüler nach bestimmten, mit der Schulleistung in Zusammenhang stehenden Merkmalen sortiert. In der überwiegenden Mehrheit der Listen wurden das Geschlecht und die Klassenzugehörigkeit als Stratifizierungsvariablen verwendet. Im Gegensatz zur ÜGK 2016 wurden jedoch keine expliziten Strata innerhalb der Schulen gebildet, da die Schülerinnen und Schüler des 8. Schuljahres nicht in getrennten Schulprogrammen unterrichten werden.

4.4 Ziehung der Schülerinnen und Schüler

Die Ziehung der Schülerinnen und Schüler wurde mithilfe der Software IBM SPSS Complex Samples 20 durchgeführt. Dazu wurden sämtliche Listen wählbarer Schülerinnen und Schüler separat eingelesen, die Listen nach den in Abschnitt 4.3 beschriebenen Merkmalen sortiert bzw. gruppiert und die Anzahl zu ziehender Schülerinnen und Schüler pro Stratum festgesetzt, bevor die entsprechende Stichprobe gezogen wurde.

Die Systematik der Methode der Stichprobenziehung auf Schülerebene ist mit der in Abschnitt 3.2 dargestellten Ziehung – mit Ausnahme des *PPS*-Verfahrens – vergleichbar: Mithilfe eines zuvor definierten Samplingintervalls wurde innerhalb jedes Stratums die Liste der wählbaren Schülerinnen und Schüler «durchgezählt» und die entsprechenden «Treffer» zur Studienteilnahme aufgeboten. Die daraus resultierende Liste der gewählten Schülerinnen und Schüler wurde anschliessend an die entsprechenden Schulen zur Information und Kontrolle geschickt.

5 Stichprobengewichte

Die zur Berechnung der Stichprobengewichte verwendeten Methoden orientieren sich an international etablierten Schulleistungsstudien wie PISA, *Trends in International Mathematics and Science Study* (TIMSS) oder *Progress in International Reading Literacy Studies (*PIRLS*)*. Die entsprechenden statistischen Theorien können beispielsweise bei Cochran (1977) oder Lohr (2010) nachgelesen werden.

Für die Analysen der im Rahmen der ÜGK gewonnenen Daten, die adäquate Berechnung der Stichprobenvarianz sowie um valide Schlussfolgerungen in Bezug auf die untersuchte Population treffen zu können, ist der Einbezug von Stichprobengewichten zwingend erforderlich. Für sämtliche an der ÜGK 2017 teilnehmenden Schülerinnen und Schüler wurden individuelle Schülergewichte (vgl. 5.5) sowie weitere Variablen berechnet, welche die Berechnung von Standardfehlern, Signifikanztests oder Konfidenzintervallen erlauben. Die Auswahlwahrscheinlichkeiten der teilnehmenden Schülerinnen und Schüler unterscheiden sich zum Teil beträchtlich. Da die einzelnen Schülerinnen und Schüler deshalb einen jeweils unterschiedlich grossen Anteil der Population repräsentieren, ist es unbedingt notwendig, die Gewichte in sämtlichen Analysen zu berücksichtigen.

Aufgrund der unterschiedlichen Auswahlsätze zwischen den Kantonen variieren die Gewichte auf nationaler Ebene relativ stark. Innerhalb der Kantone ist diese Varianz deutlich kleiner. Nebst den bereits in Abschnitt 4.2 gelisteten Gründen für variable Stichprobengewichte sind die schliesslich gültigen Stichprobengewichte auch aufgrund von *Non-Response* nicht für jede Schülerin und jeden Schüler innerhalb eines expliziten Stratums identisch:

- Gewählte Schülerinnen und Schüler, die nicht ausgeschlossen worden waren und aufgrund von Verweigerung, Krankheit oder sonstigen Gründen nicht an der Erhebung teilnahmen (*Non-Response* auf Ebene der Schülerinnen und Schüler), wurden mit Gewichtsanpassungen seitens der teilnehmenden Schülerinnen und Schüler kompensiert (*Non-Response-Adjustment; NRA*).
- Gezogene Schulen, die eine Studienteilnahme verweigerten und für die in nützlicher Frist keine Ersatzschulen gefunden werden konnten (*Non-Response* auf Schulebene), wurden mit Gewichtsanpassungen seitens der teilnehmenden Schulen kompensiert.

Das Gewicht des Schülers j oder der Schülerin j in der Schule i setzt sich aus dem Basisgewicht der Schule (w_{1i}), dem Basisgewicht innerhalb der Schule (w_{2ij}) und zwei Korrekturfaktoren zusammen (f_{1i} und f_{2ij}).

$$W_{ij} = w_{1i}w_{2i}f_{1i}f_{2ij}$$

Die einzelnen Komponenten dieses Produkts werden in den folgenden Abschnitten näher erläutert.

5.1 Basisgewicht der Schule

Für Schulen in *Kantonen mit einstufigen Stichprobenverfahren* entspricht das Basisgewicht der Schule stets $w_{1i} = 1$, da alle Schulen an der Erhebung teilgenommen haben und die einzelnen Schulen grundsätzlich nicht andere Schulen in der Population repräsentieren. In *Kantonen mit zweistufigen Stichprobenverfahren* entspricht das Basisgewicht der Schule $w_{1i} = SI / MOS$ (vgl. 3.3 – 3.5), sofern $MOS < SI$. Wenn $MOS \geq SI$ gilt, dann beträgt das Basisgewicht der Schule $w_{1i} = 1$, da diese mit einer Wahrscheinlichkeit von $p = 1$ in die Stichprobe aufgenommen wurde. Beispielsweise würde eine Schule mit $MOS = 50$ in einem Kanton mit $MOS_{tot} = 5'000$ und einem Stichprobenumfang von 40 Schulen das Basisgewicht $w_{1i} = 2.5$ erhalten ($SI / MOS = MOS_{tot} / n / MOS = 5'000 / 40 / 50$) und würde dementsprechend 2.5 Schulen im entsprechenden expliziten Stratum repräsentieren.

5.2 Basisgewicht innerhalb der Schule

Das Basisgewicht innerhalb der Schule steht für die Anzahl Schülerinnen und Schüler, die ein gezogener Schüler oder eine gezogene Schülerin in seiner bzw. ihrer Schule repräsentiert. Das Basisgewicht innerhalb der Schule ergibt sich aus dem Kehrwert der Auswahlwahrscheinlichkeit und war für alle Schülerinnen und Schüler einer Schule identisch.

5.3 Non-Response-Korrekturfaktor auf Schulebene

Ein kleiner Teil der Schulen verweigerte die Teilnahme an der ÜGK 2017 oder konnte aufgrund technischer Probleme bzw. ungenügender Infrastruktur nicht getestet werden. In Fällen rechtzeitig kommunizierter Teilnahmeverweigerungen oder technischer Probleme wurden Ersatzschulen (vgl. 3.4) kontaktiert. In einigen Fällen war es aus zeitlichen Gründen nicht mehr möglich, Ersatzschulen zur Teilnahme aufzubieten. Derartige Ausfälle wurden mittels *NRA* kompensiert.

Abgesehen von der Schulgrösse und der Trägerschaft, handelte es sich bei der ÜGK 2017 – im Gegensatz zur ÜGK 2016, bei welcher Schulen mit unterschiedlichen, leistungsabhängigen Schulprogrammen berücksichtigt werden mussten – um eine sehr homogene Schulstichprobe. Aus diesem Grund wurden Schulen im Falle einer Teilnahmeverweigerung auf Schulebene und einer fehlenden Ersatzschule mit Gewichtsanpassungen der übrigen Schulen desselben Kantons kompensiert. Dies bedeutet, dass jeder Kanton eine sogenannte *NRA*-Zelle (*Non-Response-Adjustment cell*) repräsentierte.

Innerhalb der Kantone wurde der Korrekturfaktor gemäss folgender Formel berechnet:

$$f_{1i} = \frac{\sum_{k \in A(i)} w_{1k} n(k)}{\sum_{k \in P(i)} w_{1k} n(k)}$$

Die Summe im Zähler bezieht sich auf die mit dem Schulgewicht gewichtete Anzahl Schülerinnen und Schüler *sämtlicher gezogener* (*A*) Schulen und repräsentiert somit die Population der Schülerinnen und Schüler im jeweiligen Kanton. Der Nenner hingegen beinhaltet die gewichtete Summe der Schülerinnen und Schüler aus *teilnehmenden* Schulen (*P*). Der aus dieser Formel resultierende Korrekturfaktor wurde mit jedem Schulgewicht multipliziert, damit die gesamte Population des jeweiligen Kantons durch die teilnehmenden Schulen repräsentiert wurde.

5.4 Non-Response-Korrekturfaktor auf Ebene der Schülerinnen und Schüler

Um nichtteilnehmende Schülerinnen und Schüler zu kompensieren, wurden Gruppen ähnlicher Schülerinnen und Schüler gebildet. Dazu wurden Schülerinnen und Schüler mit demselben Geschlecht der gleichen *NRA*-Zelle zugeordnet. Zur Vermeidung allzu grosser Gewichtskorrekturen wurden ausschliesslich Zellen mit mindestens 15 teilnehmenden Schülerinnen und Schülern gebildet. Dies führte dazu, dass die entsprechenden Zellen nur in Ausnahmefällen innerhalb einer Schule gebildet werden konnten.[6] Wurden Schülerinnen und Schüler aus verschiedenen Schulen gruppiert, wurde darauf geachtet, dass die entsprechenden Schulen möglichst gleich gross waren. Innerhalb der *NRA*-Zellen wurden Korrekturfaktoren gemäss folgender Formel berechnet.

[6] Damit einer *NRA*-Zelle mindestens 15 teilnehmende Knaben oder Mädchen zugeordnet werden konnten, waren schulübergreifende Zellenbildungen nicht zu vermeiden. Nur in Schulen, in denen eine verhältnismässig grosse Anzahl Schülerinnen und Schüler getestet wurde (*Kantone mit einstufigen Stichprobenverfahren*), war die Bildung von *NRA*-Zellen innerhalb einer einzelnen Schule möglich.

$$f_{2i} = \frac{\sum_{k \in A(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in P(i)} f_{1i} w_{1i} w_{2ik}}$$

Für jeden Schüler und jede Schülerin wurde das Produkt aus Basisgewicht der Schule (korrigiert für *Non-Response* auf Schulebene; $f_{1i}\ w_{1i}$) und Basisgewicht innerhalb der Schule (w_{2ik}) gebildet. Die Summe im Zähler der Formel enthält die Gewichte sämtlicher gezogener Schülerinnen und Schüler mit Ausnahme der ausgeschlossenen Fälle (z.B. kognitive Beeinträchtigung, vgl. 2.2). Aus der Division dieser Summe mit der Summe der Gewichte der teilnehmenden Schülerinnen und Schüler ergeben sich die Korrekturfaktoren für jede *NRA*-Zelle. Mit anderen Worten: Die Gewichte der teilnehmenden Schülerinnen und Schüler wurden dann erhöht, wenn hinsichtlich Geschlecht sowie Schulgrösse «ähnliche» Schülerinnen und Schüler abwesend waren. Ausgeschlossene Schülerinnen und Schüler wurden nicht mittels *NRA* kompensiert.

6 Berechnung der Stichprobenvarianz

Bei den im vorliegenden Bericht vorgestellten Stichprobenverfahren handelt es sich jeweils um eine Zufallsauswahl. Dadurch besitzt die Stichprobe eine Wahrscheinlichkeitsverteilung und ermöglicht die Anwendung inferentieller Statistik (z.B. Signifikanztests, Schätzer mit Konfidenzintervallen usw.; vgl. von der Lippe & Kladroba, 2002). Die Zufallskomponente führt dazu, dass die Anteile der Schülerinnen und Schüler, welche die Grundkompetenzen erreichen, – oder beliebige auf dem ÜGK-Datensatz beruhende Schätzer – von der Stichprobe abhängig sind und somit je nach Auswahl von Schulen oder Schülerinnen und Schülern variieren können. Die Stichprobenvarianz beziffert, inwiefern sich die Ergebnisse ändern würden, wenn die Grundkompetenzen auf Basis anderer Schülerinnen und Schüler der Population überprüft worden wären.

Zur Berechnung der Stichprobenvarianz ist es zwingend erforderlich, das komplexe Stichprobendesign sowie die entsprechende Gewichtung zu berücksichtigen. Eine Übersicht entsprechender Methoden bieten beispielsweise Wolter (1985) oder Lee, Forthofer und Lorimor (1989). Sowohl aus praktischen als auch historischen Gründen haben sich im Rahmen nationaler (z.B. *National Assessment of Educational Progress; NAEP*) oder internationaler (z.B. *PISA, TIMSS* oder *PIRLS*) *Large-Scale-Assessments* die Replikationsverfahren als Standard zur Schätzung der Stichprobenvarianz etabliert (vgl. Rust, 2014). Dementsprechend wird in Abschnitt 6.1 ein bestimmtes Replikationsverfahren, auf welchem die im Datensatz enthaltenen *Replicate Weights* beruhen, näher vorgestellt. Dennoch kann für die Varianzschätzung auch auf Linearisierungsverfahren (Demnati & Rao, 2004) zurückgegriffen werden. Eine kurze Gegenüberstellung von Replikations- und Linearisierungsverfahren – auch in Abhängigkeit von statistischer Analysemethode und Art des Schätzers – bietet Valliant (2007).

Im Rahmen von Analysen, die sich auf Leistungsvariablen beziehen (z.B. Anteile der Schülerinnen und Schüler pro Kanton, die den Grundkompetenzen nicht genügen), sollte zusätzlich stets der mit den Tests verbundene Messfehler berichtet werden. Dieser kann mittels der *Plausible Values* (vgl. Angelone & Keller, 2019) der skalierten Leistungswerte berechnet werden. Der Standardfehler eines entsprechenden Schätzers setzt sich dementsprechend aus Stichprobenvarianz und Messfehler zusammen. Da Berechnungen mit *Plausible Values* unter gleichzeitiger Verwendung von Varianzschätzverfahren komplexe und rechenintensive Analysen darstellen, ist im Anhang C ein kurzes Auswertungsbeispiel mithilfe eines *R*-Programmpakets dargestellt.

6.1 Die «Balanced Repeated Replication»-Methode

Die Grundidee von Replikationsverfahren besteht darin, die Stichprobenvarianz eines Ergebnisses zu berechnen, indem dieses mehrmals – mit einer jeweils unterschiedlichen Gewichtung einzelner Studienteilnehmer – geschätzt wird. Die

Stichprobenvarianz wird aus der Variabilität des mehrmals berechneten Ergebnisses abgeleitet. Die bei der ÜGK 2017 verwendete bzw. im ÜGK-Datensatz integrierte Methode zur Schätzung der Stichprobenvarianz nennt sich *Balanced Repeated Replication* (*BRR*; vgl. Rust, 1985; Rust & Rao, 1996). Analog zur Vorgehensweise bei PISA wurde die Variante des Verfahrens gewählt, die als *Fay's Methode* bekannt ist (vgl. Judkins, 1990). Der Datensatz ÜGK 2017 enthält 120 zusätzliche Gewichtsvariablen bzw. unterschiedliche Kombinationen von Schülergewichten (*Replicate Weights*), die jeweils neue Varianten der Stichprobe darstellen und sich durch eine veränderte Gewichtung der gezogenen Schülerinnen und Schüler von der anfänglichen Stichprobe unterscheiden. Die Stichprobenvarianz eines Ergebnisses wird beruhend auf der Variabilität der entsprechenden Werte zwischen den 120 neugebildeten Varianten der Stichprobe geschätzt. Konkret bedeutet dies, dass die Stichprobenvarianz beliebiger – mit der ÜGK 2017 berechneter – Schätzer mittels der folgenden Formel berechnet werden kann:

$$\mathrm{V}_{BRR}(X^*) = \frac{1}{30}\sum_{t=1}^{120}\{(X_t^* - X^*)^2\}$$

Dabei entspricht X^*_t der Schätzung des interessierenden Merkmals, beruhend auf der wiederholten Über- und Untergewichtung mithilfe der neu erstellten Varianten der Stichprobe, und X^* der Schätzung, basierend auf der anfänglichen Stichprobe (Ausgangsgewichtung). Die Erstellung der 120 alternativen Gewichtungen folgt einer bestimmten Methode, die im folgenden Abschnitt erläutert wird.

6.2 Berechnung der «Replicate Weights»

Die Berechnung der *Replicate Weights* beinhaltet die Bildung von Paaren aus Stichprobeneinheiten, die in der Summe stets dasselbe Gewicht behalten, bei denen die einzelnen Stichprobeneinheiten allerdings unterschiedlich stark gewichtet werden. Im Rahmen der ÜGK 2017 wurden dazu teilnehmende Schulen (in *Kantonen mit zweistufigen Stichprobenverfahren*) bzw. Schülerinnen und Schüler (in *Kantonen mit einstufigen Stichprobenverfahren*) auf Grundlage der Stratifizierungsvariablen gepaart, bevor innerhalb jedes Paares eine Einheit stärker und eine Einheit schwächer gewichtet wurde. Dazu wurden die entsprechenden Gewichte innerhalb jedes Paares mit den Faktoren 0.5 bzw. 1.5 multipliziert (*Fay's Variante* der *BRR*-Methode; vgl. Judkins, 1990), sodass die Summe der Gewichte innerhalb jedes Schulpaares unverändert blieb. Dieses Vorgehen wurde 120-mal wiederholt, wobei jeweils neue Kombinationen aus über- und untergewichteten Einheiten entstanden. Im Detail enthielt die Berechnung der *Replicate Weights* für die ÜGK 2017 die folgenden Schritte:

- In *Kantonen mit zweistufigen Stichprobenverfahren* wurden auf Basis von expliziter und impliziter Stratifizierung Schulpaare gebildet. Die Paare (bei einer ungeraden Anzahl Einheiten wurde ein *Triple* gebildet) repräsentieren sogenannte Varianzstrata, die lediglich zur Schätzung der Stichprobenvarianz dienen. Diese wurden in 116 grössere Varianzstrata zusammengefasst.
- Innerhalb der Stichprobenhälften wurde stets das Basisgewicht einer Schule (w_{1i}, vgl. 5.1) mit dem Faktor 1.5 übergewichtet, während dasjenige der anderen Schule mit dem Faktor 0.5 untergewichtet wurde. Bei *Triples* wurden leicht angepasste Über- und Untergewichtungsfaktoren verwendet (vgl. hierzu OECD, 2017). Die Verteilung dieser Gewichtungsfaktoren folgte einer Hadamardmatrix (orthogonale Matrix, vgl. Lee, Forthofer & Lorimor, 1989, S. 30), die in 120 unterschiedlichen Kombinationen von Stichprobenhälften resultierte.
- Analog zu den *Replicate Weights* auf Schulebene wurden für *Kantone mit einstufigen Stichprobenverfahren* Paare auf Schülerebene gebildet. Die auf Basis der impliziten Stratifizierung gebildeten Paare wurden anschliessend zu 116 grösseren Gruppen zusammengefasst.
- Die Basisgewichte der Schülerinnen und Schüler wurden in unterschiedlicher Kombination wiederum hälftig um den Faktor 1.5 erhöht bzw. um den Faktor 0.5 reduziert. Hieraus ergaben sich erneut 120 *Replicate Weights* (für *Triples* vgl. OECD, 2017).
- Die 120 neu gebildeten Gewichte *(Replicate Weights)* wurden schliesslich gemäss den in Abschnitt 5.3 und 5.4 beschriebenen Korrekturen für *Non-Response* auf Schul- sowie Schülerebene separat angepasst.

7 Literatur

Angelone, D. & Keller, F. (2019). *ÜGK 2017 Schulsprache und erste Fremdsprache. Technische Dokumentation zur Testentwicklung und Skalierung.* Aarau: Geschäftsstelle der Aufgabendatenbank EDK (ADB).

Cochran, W. G. (1977). *Sampling Techniques.* New York: John Wiley and Sons.

Demnati, A. & Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology, 30*, 17–26.

Judkins, D. R. (1990). Fay's Method of Variance Estimation. *Journal of Official Statistics, 6*(3), 223–239.

Kauermann, G. & Küchenhoff H. (2011). *Stichproben*. Heidelberg: Springer.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistics, 8*, 183–200.

Kish, L. (1995). Methods for Design Effects. *Journal of Official Statistics, 11*, 55–77.

Le, T., Brick, M. & Kalton, G. (2002). Decomposing Design Effects. In *JSM Proceedings, Survey Research Methods Section* (S. 2007–2012). Alexandria, VA: American Statistical Association.

Lee, S. L., Forthofer, R. N. & Lorimor, R. J. (1989). *Analyzing Complex Survey Data* (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 07-064). Newbury Park, CA: Sage.

Lehtonen, R. & Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Survey.* Chichester: John Wiley and Sons.

Liu, J., Iannacchione, V. & Byron, M. (2002). Decomposing Design Effects for Stratified Sampling. In *JSM Proceedings, Survey Research Methods Section* (S. 2124–2126). Alexandria, VA: American Statistical Association.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Boston, MA: Brooks/Cole.

Meinck, S. (2015). Computing Sampling Weights in Large-scale Assessments in Education. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach. Verfügbar unter: http://surveyinsights.org/?p=5353 [2.05.2019].

OECD (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing.

Pham, G. (2019). *ÜGK 2017 – Technical report: Student questionnaire data*. St. Gallen: Pädagogische Hochschule St. Gallen.

Robitzsch, A. & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 3.2-25*. Verfügbar unter: https://CRAN.R-project.org/package=BIFIEsurvey [3.05.2019].

Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics, 1*, 381–397.

Rust, K. (2014). Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis (S. 117–153). Boca Raton, FL: CRC Press.

Rust, K. & Rao, J. N. K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Survey Methods in Medical Research, 5*, 283–310.

Sacchi, S. & Oesch, D. *ÜGK 2016: Assessment of mathematics skills. Documentation of questionnaire-based scales.* Bern: TREE, Universität Bern.

Valliant, R. (2007). An Overview of the Pros and Cons of Linearization versus Replication in Establishment Surveys. *Papers presented at the ICES-III*, 929-940.

von der Lippe, P. & Kladroba, A. (2002). Repräsentativität von Stichproben. *Marketing ZFP – Journal of Research and Management, 24*, 139–144.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer.

8 Anhang A: Kennzahlen zur Stichprobenziehung

8.1 Ausschluss- und Ausschöpfungsquoten

Wie in Kapitel 2 beschrieben, galten bestimmte Schülerinnen und Schüler der *erwünschten Population* als «nicht erreichbar» und hatten eine Wahrscheinlichkeit von $p = 0$ in die Stichprobe aufgenommen zu werden. Dies betraf einzelne Schülerinnen und Schüler aus Regelschulen, die von den zuständigen Lehrpersonen ausgeschlossen wurden, sowie sämtliche in Sonderschulen unterrichteten Schülerinnen und Schüler. Die entsprechenden Ausschlussquoten werden in der Tabelle A.1 getrennt nach Kanton aufgeführt. Die geschätzten Ausschöpfungsquoten widerspiegeln den Anteil der *erwünschten Population*, der durch die *ÜGK-Population* abgedeckt werden konnte. Da über die ausgeschlossenen Schülerinnen und Schüler anhand der ÜGK 2017 keine Aussagen getroffen werden können, sollten die Ausschöpfungsquoten als Interpretationshilfe bei kantonalen Leistungsvergleichen herangezogen werden.

In zahlreichen Kantonen können in Sonderschulen unterrichtete Schülerinnen und Schüler nicht einem bestimmten Schuljahr zugeordnet werden. Die kantonalen Anteile in Sonderschulen unterrichteter Schülerinnen und Schüler wurden deshalb auf Basis der SDL aus dem Schuljahr 2016/17 geschätzt, indem in Sonderschulen unterrichtete 12-Jährige ins Verhältnis zur gesamten 12-jährigen Schülerschaft eines Kantons gesetzt wurden.

Tabelle A.1: Anzahl bzw. Anteile der bei der ÜGK 2017 ausgeschlossenen Schülerinnen und Schüler sowie geschätzte Ausschöpfungsquoten getrennt nach Kanton

	In Regelschulen ausgeschlossene SuS getrennt nach Ausschlussgrund (absolute Anzahl)										In Regelschulen ausgeschlossene SuS (Total)		Ausgeschlossene SuS der erwünschten Population in %		Geschätzte Ausschöpfungsquote der erwünschten Population in %
Kanton	*a)*	*b)*	*c)*	*d)*	*e)*	*f)*	*g)*	*h)*	*i)*	*j)*	ungew.	gew.	Regel-schulen	Sonder-schulen	
AG	4	2	0	0	0	0	0	0	0	1	7	47	0.7	2.7	96.6
AI	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	100.0
AR	0	0	0	0	0	0	0	0	0	1	1	1	0.2	4.3	95.5
BE_d	7	5	3	3	1	0	0	0	0	1	20	170	2.0	1.8	96.2
BE_f	1	1	1	0	0	0	0	0	0	0	3	5	0.6	0.9	98.5
BL	4	0	2	0	0	0	0	0	1	0	7	22	0.9	1.4	97.7
BS	1	2	1	1	0	0	0	0	4	0	9	16	1.1	0.5	98.4
FR_d	1	4	0	0	0	0	0	0	1	1	7	10	1.3	0.0	98.7
FR_f	5	10	1	0	3	0	1	0	4	2	26	70	2.6	0.0	97.4
GE	4	7	1	0	0	0	0	0	2	1	15	70	1.5	1.3	97.2
GL	0	0	0	0	0	0	0	0	0	0	0	0	0.0	1.5	98.5
GR	2	1	2	0	2	0	0	1	2	1	11	16	1.2	1.5	97.3
JU	0	1	0	0	0	0	0	0	1	0	2	3	0.4	1.4	98.3
LU	3	3	4	2	1	0	0	0	0	0	13	53	1.4	1.6	97.1
NE	0	1	0	0	1	0	0	0	11	0	13	33	1.7	1.4	96.8
NW	2	2	1	0	0	0	0	0	1	3	9	11	3.1	0.8	96.0
OW	1	0	0	0	0	0	0	0	0	0	1	1	0.3	0.6	99.2
SG	0	3	0	0	0	1	0	0	0	0	4	18	0.4	2.2	97.4
SH	1	1	0	2	0	0	0	0	0	0	4	5	0.7	2.4	96.9
SO	6	7	2	0	0	0	0	0	0	2	17	36	1.5	3.2	95.3
SZ	2	2	0	1	0	0	0	0	2	0	7	15	1.1	0.5	98.3
TG	4	2	0	0	0	0	0	0	0	0	6	17	0.6	0.3	99.1
TI	7	1	0	3	0	0	0	0	2	0	13	56	1.8	2.1	96.1
UR	1	0	0	0	0	0	0	0	0	0	1	1	0.3	0.6	99.1
VD	0	4	0	0	0	2	0	0	4	1	11	88	1.1	2.3	96.6
VS_d	5	2	0	0	1	0	0	0	2	0	10	15	2.0	0.8	97.3
VS_f	11	8	4	2	0	0	0	0	3	2	30	81	3.2	0.8	96.0
ZG	2	3	3	2	0	0	0	0	2	1	13	29	2.6	3.5	93.8
ZH	9	7	2	2	1	0	0	0	1	0	22	163	1.2	1.8	97.0
CH	83	**79**	**27**	**18**	**10**	**3**	**1**	**1**	**43**	**17**	**277**	**1'052**	**1.3**	**1.6**	**97.1**

Ausschlussgründe: a) Geringe Kenntnisse der Testsprache; *b)* Lernbehinderung; *c)* Kognitive Beeinträchtigung; *d)* Verhaltensbehinderung; *e)* Sprachbehinderung; *f)* Sehbehinderung; *g)* Hörbehinderung; *h)* Körperbehinderung; *i)* Mehrfachbehinderung; *j)* Andere Gründe.

Anmerkungen: In Sonderschulen unterrichtete Schülerinnen und Schüler können in den meisten Fällen nicht einer bestimmten Klassenstufe zugeordnet werden. Die hier dargestellten Schätzungen beruhen auf Anteilen von SuS eines bestimmten Jahrgangs.

8.2 Rücklaufquoten auf Schulebene

Ein kleiner Teil der gezogenen bzw. zur Erhebung aufgebotenen Schulen hat die Teilnahme an der ÜGK 2017 verweigert. In *Kantonen mit zweistufigen Stichprobenverfahren* wurden deshalb Ersatzschulen gezogen (vgl. 3.3), die teilweise für verweigernde Schulen eingesprungen sind. In der Mehrzahl der Kantone funktionierte das Ersetzen verweigernder Schulen derart gut, dass ein Rücklaufquote von 100 Prozent erreicht werden konnte. Die in Tabelle A.2 dargestellten Rücklaufquoten wurden mit der Anzahl unterrichteter Schülerinnen und Schüler der getesteten Schulen gewichtet.

Tabelle A.2: Anzahl untersuchter Schulen und mit Schülerbestand gewichtete Rücklaufquoten auf Schulebene getrennt nach Kanton

Kanton	Anzahl untersuchter Schulen (ohne Ersatzschulen)	Anzahl untersuchter Ersatzschulen	Mit Schülerbestand gewichtete Rücklaufquote
AG	53	0	100.0%
AI	10	-	100.0%
AR	31	-	100.0%
BE_d	59	4	100.0%
BE_f	32	-	100.0%
BL	55	3	100.0%
BS	32	-	100.0%
FR_d	30	-	100.0%
FR_f	54	0	100.0%
GE	49	0	100.0%
GL	18	-	99.7%
GR	62	1	99.0%
JU	42	-	96.4%
LU	56	0	100.0%
NE	31	0	100.0%
NW	16	-	100.0%
OW	12	-	100.0%
SG	57	7	98.9%
SH	35	-	96.6%
SO	57	2	98.0%
SZ	35	2	100.0%
TG	60	0	100.0%
TI	24	0	100.0%
UR	20	-	100.0%
VD	50	0	100.0%
VS_d	40	-	100.0%
VS_f	55	0	98.0%
ZG	40	-	100.0%
ZH	88	2	99.3%
CH	**1'203**	**21**	**99.2%**

Anmerkung: In *Kantonen mit einstufigen Stichprobenverfahren* wurden alle Schulen zur Erhebung aufgeboten, weshalb keine Ersatzschulen bestimmt werden konnten (vgl. «-» in der Spalte «Anzahl getesteter Ersatzschulen»).

8.3 Rücklaufquoten auf Schülerebene

Schülerinnen und Schüler, die nicht mindestens eine gültige Antwort in den Testaufgaben oder im Fragebogen gegeben haben und gleichzeitig nicht von der Erhebung ausgeschlossen wurden (vgl. die Tabellen A.1 und A.2), galten als abwesend. Für die grosse Mehrheit der abwesenden Schülerinnen und Schüler wurde von Seiten der Schulen kein Abwesenheitsgrund kommuniziert. Abwesenheiten aufgrund von technischen Problemen oder Verweigerung der Eltern sind äusserst selten vorgekommen. Um Rücklaufquoten auf Schülerebene zu berechnen, wurden die Summen von getesteten sowie von allen zur Erhebung aufgebotenen Schülerinnen und Schülern (ohne Ausschlüsse) ins Verhältnis gesetzt. Entsprechende ungewichtete und gewichtete (Gewichtung mit w_{1i}, w_{2ij} sowie f_{1i}; vgl. Kapitel 5) Rücklaufquoten können Tabelle A.3 entnommen werden.

Tabelle A.3: Ungewichtete und gewichtete Anzahl erhobener, abwesender Schülerinnen und Schüler sowie entsprechende Rücklaufquoten

	Ungewichtete Anzahl Schülerinnen und Schüler		Gewichtete Anzahl Schülerinnen und Schüler		Rücklaufquoten in %	
Kanton	erhoben	abwesend	erhoben	abwesend	ungewichtet	gewichtet
AG	906	27	5'969	173	97.1	97.2
AI	139	2	139	2	98.6	98.6
AR	441	14	454	14	96.9	97.0
BE_d	940	44	7'825	348	95.5	95.7
BE_f	527	31	755	40	94.4	95.0
BL	865	35	2'292	93	96.1	96.1
BS	628	39	1'334	83	94.2	94.1
FR_d	513	14	740	17	97.3	97.8
FR_f	955	18	2'597	48	98.2	98.2
GE	892	27	4'470	139	97.1	97.0
GL	254	6	330	9	97.7	97.3
GR	820	29	1'270	46	96.6	96.5
JU	595	20	771	28	96.7	96.5
LU	940	15	3'741	57	98.4	98.5
NE	643	22	1'773	62	96.7	96.6
NW	273	6	333	7	97.8	97.9
OW	246	5	345	6	98.0	98.3
SG	951	29	4'455	130	97.0	97.2
SH	534	16	630	16	97.1	97.5
SO	912	30	2'203	72	96.8	96.8
SZ	654	18	1'277	34	97.3	97.4
TG	927	37	2'691	108	96.2	96.1
TI	744	24	3'048	95	96.9	97.0
UR	290	7	319	8	97.6	97.6
VD	928	33	7'718	264	96.6	96.7
VS_d	572	12	730	15	97.9	98.0
VS_f	944	16	2'383	42	98.3	98.3
ZG	577	37	978	56	94.0	94.6
ZH	1'567	74	12'771	640	95.5	95.2
CH	**20'177**	**687**	**74'341**	**2'652**	**96.7**	**96.6**

9 Anhang B: Zusatzinformationen zu Schulstichproben

9.1 Umgang mit kleinen und sehr kleinen Schulen

Enthielten Strata Schulen mit weniger als 20 Schülerinnen und Schülern, bestand das Risiko, dass die Anzahl gezogener Schülerinnen und Schüler dem erwünschten Stichprobenumfang nicht genügt. Aus diesem Grund wurden vor der Schulziehung die Listen wählbarer Schulen mithilfe der folgenden Schritte analysiert (vgl. auch 3.4 sowie OECD, 2017, S. 77):

- Innerhalb jedes expliziten Stratums wurde der gesamthafte Schülerbestand prozentual auf *sehr kleine* Schulen (K1; Schülerbestand < 3), *kleine Schulen* (K2; Schülerbestand ≥ 3 und < 10), *mittelgrosse* Schulen (M; Schülerbestand ≥ 10 und < 20) und *grosse* Schulen (G; Schülerbestand ≥ 20) aufgeteilt, sodass K1 + K2 + M + G = 1 galt.
- Wenn K1 + K2 ≤ 0.01 war und somit über 1 Prozent des Schülerbestands im Stratum ausmachte, wurden *kleine* sowie *sehr kleine* Schulen unterrepräsentiert (*TCS* angepasst) und die Anzahl zu ziehender Schulen angehoben.
- Wenn K1 + K2 < 0.01 und M ≥ 0.04 galt, dann wurde lediglich die Anzahl zu ziehender Schulen angehoben.
- Wenn K1 + K2 < 0.01 und M < 0.04 galt, dann wurden keine Anpassungen vorgenommen.

Falls die Anzahl zu ziehender Schulen angehoben werden musste, wurden neue Schulstichprobenumfänge beruhend auf den folgenden Formeln berechnet:

- Berechnung des Faktors L = 1 + 3 (K1) / 4 + (K2) / 2.
- Berechnung der mittleren Schulgrösse für *mittelgrosse* Schulen (MENR), *kleine* Schulen (K2ENR) und *sehr kleine* Schulen (K1ENR).
- Der minimale Schulstichprobenumfang für *grosse* Schulen entsprach der ursprünglichen Anzahl zu ziehender Schulen multipliziert mit G und L.
- Der minimale Schulstichprobenumfang für *mittelgrosse* Schulen entsprach (N/2 · M · L) / K2ENR, wobei N den Schülerbestand in *mittelgrossen* Schulen repräsentiert.
- Der minimale Schulstichprobenumfang für *sehr kleine* Schulen entsprach (N/4 · K1 · L) / K1ENR, wobei N den Schülerbestand in *sehr kleinen* Schulen repräsentiert.

10 Anhang C: Auswertungshinweise

In der Folge wird am Beispiel der Berechnung kantonaler Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen erreichen, der adäquate Einbezug von Schülergewichten, *Plausible Values* und *Replicate Weights* mit dem *R*-Paket *BIFIE-Survey* (Robitzsch & Oberwimmer, 2019) kurz illustriert. Es soll hervorgehoben werden, dass sich – bedingt durch die multiple Imputation zahlreichen Variablen im Rahmen der ÜGK 2017 (vgl. Pham, 2019) – die Datenstruktur der ÜGK 2017 deutlich von derjenigen der ÜGK 2016 unterscheidet.

Nach dem Start von *R*, ist es in einem ersten Schritt notwendig, das Paket *BIFIE-Survey* zu installieren und zu laden.

```
install.packages("BIFIEsurvey")
library(BIFIEsurvey)
```

Anschliessend bietet es sich an, einen lokalen Ordner (Beispielpfad: C:/UEGK_Daten_2017) zu definieren, in welchem sich die 20 imputierten Datensätze befinden, bevor sämtliche Dateien im *Rdata*-Format einer Liste hinzugefügt werden.

```
Folder <- "C:/UEGK_Daten_2017"
files <- list.files(path = folder,pattern = "UEGK_2017__IMP-
DATA(.)*.Rdata",full.names = T)
```

Mithilfe eines Schlaufenbefehls können die 20 Datensätze einem R-Objekt hinzugefügt werden.

```
datlist <- list()
for(ii in 1:20){
  lf <- files[ii]
  load.Rdata(lf,"dat")
  datlist[[ii]] <- dat
  print(paste0("Done ",ii))
}
```

Zur korrekten Berechnung von Stichprobenfehlern müssen die *Replicate Weights* definiert werden. Da sämtliche *Replicate Weights* das Präfix *smp_w_nrasturw* im Variablennamen tragen, können diese mithilfe des *grep*-Befehls in *R* identifiziert werden.

```
reps.col <- grep("smp_w_nrasturw", names(datlist[[1]]))
```

Sämtliche Berechnungen in *BIFIE-Survey* greifen auf *BIFIE.dat*-Objekte zu, die zunächst erstellt werden müssen. Dabei werden die Datensätze eingelesen und Schülergewichte sowie *Replicate Weights* definiert.

```
uegk17 <- BIFIE.data(data.list=datlist,
wgt=datlist[[1]]$smp_w_nrastubw,
wgtrep=datlist[[1]][,reps.col],
fayfac=(1/120*4))
```

Schliesslich können mithilfe der *Plausible Values* für eine bestimmte Skala (z.B. Lesen in der Schulsprache) und des Befehls *BIFIE.univar* die kantonalen Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen erreicht haben, geschätzt werden. Die Dokumentation bzw. Hilfefunktion in *R* zum *BIFIE-Survey*-Paket enthält wertvolle Hinweise zu zahlreichen weiteren Analysemöglichkeiten.

```
desc1 <- BIFIEsurvey::BIFIE.univar(bdat1,
vars=c("PL_PVR_SL"),
group="id_canton")
summary(desc1)
desc1$stat
```